Modélisation PINAR(p) et prévision du nombre d'admissions hospitalières

Mohamed SADOUN[1,2] & Mohamed BENTARZI[2]

mo-hamedsadoun@outlook.fr *mohamedbentarzi@yahoo.fr*

[1] *Centre for Research in Applied Economic for Development (CREAD), Algiers, Algeria*

[2] *Operational Research Department, University of USTHB, Algiers, Algeria*

Résumé. Cette communication propose une modélisation autorégressive à valeurs entières d'ordre arbitraire à coefficients périodique $PINAR(p)$, dans le but d'analyser le nombre d'arrivées aux services d'urgence hospitaliers causées par des maladies ayant un comportement saisonnier. Deux méthodes d'estimation des paramètres du modèle seront proposées, à savoir : la méthode des moindres carrées conditionnelles (MCC) et la méthode du maximum de vraisemblance conditionnelle (MVC). De plus la fonction de prédiction du modèle sera donnée en utilisant une certaine représentation de l'espérance conditionnelle. Les performances des estimateurs obtenus seront montrés via une étude de simulation intensive. Une application sur données réelles sera réalisée pour modéliser le nombre mensuel d'admissions hospitalières causées par la grippe.

Mots-clés. Processus à valeurs entières périodiquement corrélé, le processus $PINAR(p)$, estimation par (MCC), estimation par (MVC)

Abstract. This communication proposes an arbitrary-order periodic integer-valued autoregressif $PINAR(p)$ modeling, in order to analyze the number of hospital emergency department arrivals caused by diseases with seasonal behavior. Two methods of parameters estimations will be proposed, namely : the conditional least squares (CLS) and the conditional maximum likelihood (CML) methods. Moreover, the prediction function of the model will be given using some representation of the conditional expectation. The performance of the obtained estimators, will be shown via an intensive simulation study. An application on a real data set will be realized to model the number of hospital admissions per month caused by influenza.

Keywords. Periodically correlated integer-valued process, periodic $INAR(p)$ model, conditional least squares (CLS) estimation, conditional maximum likelihood (CML) estimation.

1 Introduction

Parfois, la décision d'hospitaliser ou non un patient qui arrive au service d'urgence de l'hôpital est prise en fonction non seulement de la gravité de la maladie, mais également du nombre de lits disponibles. Il est donc crucial de prévoir de manière fiable le nombre

d'arrivées futures aux services d'urgence en fonction de ce qui s'est produit dans le passé. Compte tenu du nombre d'arrivées et d'admissions aux services d'urgence, les prévisions relatives aux arrivées peuvent être utilisées pour prévoir le nombre de lits nécessaires.
Il est bien connu, de nos jours, que de nombreuses séries chronologiques à valeurs entières d'ordre médical, économique, financier et environnemental, rencontrées dans la pratique, présentent une structure d'autocorrélation périodique (par exemple, *Monthly counts of claims of short-term disability benefits* été étudiée récemment par Bourguignon *et al* (2016)). De plus, cette caractéristique périodique ne peut pas être prise en compte et décrite par des modèles de série temporelle à valeurs entières avec des paramètres invariants dans le temps. Tenant compte de ces faits et des divers avantages et propriétés intéressantes remplis par un modèle $INAR(p)$ tels que la positivité et la nature discrète des réalisations qui font de lui un processus de branchement avec immigration. Historiquement, ce modèle a été introduit et étudié par Al-Osh and Alzaid (1990) avec une structure d'autocorrélation $ARMA(p, p-1)$, mais Du and Li (1991) ont proposé une procédure différente dans laquelle la structure d'autocorrélation d'un $INAR(p)$ est identique à celle d'un $AR(p)$. Toutes ces raisons nous ont donné une motivation pour étendre cette classe de modèles $INAR$ invariante dans le temps à une classe $PINAR$ variante dans le temps de façon périodique. Les processus $(PINAR)$ ont été introduits pour modéliser les phénomènes à valeurs entières non-négatives qui évoluent dans le temps affectés par des perturbations saisonnières. La distribution d'un processus $(PINAR)$ est déterminée par deux paramètres : un vecteur de probabilités de survies périodiques et une distribution de probabilité sur des entiers non-négatives, appelée distribution d'immigration périodique. Dans le contexte de cette communication, il est important de mentionner les travaux de Moriña et al (2011) qui ont suggéré un processus $INAR(2)$ particulier avec une structure saisonnière pour analyser le nombre d'arrivées par semaine au service des urgences de l'hôpital Clinique i Provincial de Barcelone causée par la grippe.
Le reste du papier est organisé comme suit. Dans la section suivante, nous fournissons quelques notations, définitions et résultats préliminaires de base concernant le modèle autorégressif à valeurs entières à coefficients périodiques $(PINAR_S(p))$, qui nous sera nécessaire dans les sections à venir. Dans la troisième section, nous établissons d'abord les estimateurs des moindres carrées conditionnelles (MCC) et celles du maximum de vraisemblance conditionnelle (MVC), ensuite nous proposons une fonction de prédiction à court-terme pour notre modèle. La quatrième section sera consacrée à l'étude intensive de simulation ainsi que pour une application sur des données réelles concernant le nombre mensuel de personnes ayant reçu un diagnostic de grippe dans la région de Catalogne (Espagne) entre 2009 et 2016.[1]

[1]Les données ont été aimablement fournies par David Morina et Brendan McCabe

2 Définition du Modèle PINAR(p)

Brièvement, un processus périodiquement corrélé dans le sens de Gladyshev (1963) avec une période S (où S est un entier strictement positive, $S \geq 2$), $\{y_t;\ t \in \mathbb{Z}\}$, est dit satisfaisant à un modèle autorégressif à valeurs entière périodique d'ordre p, et est noté$(PINAR_S(p))$, si il est solution de l'équation suivante non-linéaire aux différences stochastiques:

$$y_t = \sum_{i=1}^{p} \varphi_{t,i} \circ y_{t-i} + \varepsilon_t,\ t \in \mathbb{Z}, \tag{2.1a}$$

où le processus à valeurs entière non-négative $\{y_t, t \in \mathbb{Z}\}$, est périodiquement corrélé, avec la période à valeurs entière positive S $(S \geq 2)$, et où le processus d'innovation $\{\varepsilon_t, t \in \mathbb{Z}\}$, représente une suite de variables aléatoires indépendantes à valeurs entières non-négative, suivant une certaine distribution de probabilité discrète appartenant à une famille de lois paramétrique de la forme $\left\{\mathbb{G}_{\underline{\alpha}_t} \left| \underline{\alpha}_t = (\alpha_{t,1}, \alpha_{t,2}, ..., \alpha_{t,q})' \in A \subset \mathbb{R}_+^q \right.\right\}$. Les vecteurs de paramètres $\underline{\varphi}_t = \left(\varphi_{t,1}, \varphi_{t,2}, ..., \varphi_{t,p}\right)'$ et $\underline{\alpha}_t$ sont périodiques, concernant t, avec une période S $(S \geq 2)$, où S est le plus petit entier positif satisfaisant la relation $\underline{\varphi}_{t+rS} = \underline{\varphi}_t$ and $\underline{\alpha}_{t+rS} = \underline{\alpha}_t$. Enfin le symbole " $\circ$ " représente, comme d'habitude, l'opérateur d'amincissement de Steutel-Van Harn (1979), qui est définit pour n'importe quel processus stochastique à valeur entière y_{t-i} et pour toute suite indépendantes de variables aléatoire de comptages à valeurs entières non-négatives $\{Y_{k,t,i}, k \in \mathbb{N}, t \in \mathbb{Z}\}$ où $P(Y_{k,t,i} = 1) = 1 - P(Y_{k,t,i} = 0) = \varphi_{t,i} \in [0,\ 1], i = 1, \ldots, p$ par

$$\varphi_{t,i} \circ y_{t-i} = \begin{cases} \sum_{k=1}^{y_{t-i}} Y_{k,t,i}, & si\ y_{t-i} > 0, \\ 0, & si\ y_{t-i} = 0. \end{cases} \tag{2.1b}$$

où les suites $\{Y_{k,t,i}\}_{k \in \mathbb{N}, t \in \mathbb{Z}, i=1,\ldots,p}$ sont indépendantes. De plus, le processus d'innovation $\{\varepsilon_t, t \in \mathbb{Z}\}$ est supposé être indépendant de y_{t-i} et $\varphi_{t,i} \circ y_{t-i}$. Il est important de mentionner qu'aucune propriété du modèle (2.1), ainsi que les deux méthodes d'estimations des paramètres du MCC et du MVC n'ont été abordé jusqu'à présent, sauf dans le cas où $p = 2$ et cela pour la méthode du MVC dans Moriña *et al* (2011). Soit $\underline{\alpha}_s = (\alpha_{s,1},\ \alpha_{s,2},\ ...,\ \alpha_{s,q})'$ le vecteur des paramètres $q-$dimensionné de la distribution du processus d'innovation, nous pouvons alors définir le vecteur des paramètres globaux du modèle (2.1) de dimension $(p+q)S$, de la sorte: $\underline{\theta} = (\underline{\theta}_1'; \underline{\theta}_2'; \ldots; \underline{\theta}_S')'$ avec $\underline{\theta}_s = \left(\underline{\varphi}_s'; \underline{\alpha}_s'\right)' \in [0,1]^p \times A \subset \mathbb{R}_+^p \times \mathbb{R}_+^q$, $s = 1, 2, \ldots, S$, où A est un sous-ensemble convexe de $\mathbb{R}_+^p$.

3 Estimation et Prévision du Modèle

Cette section est consacrée à l'estimation des paramètres du modèles $PINAR_S(p)$, tout en utilisant deux méthodes d'estimation, à savoir la méthode du (MCC) et (MVC). La deuxième partie de la section sera réservée à l'étude de la fonction prédictive de notre modèle. Nous faisons l'hypothèse que $\{\varepsilon_t, t \in \mathbb{Z}\}$ est une suite $i.i.d$ suivant une distribution paramétrique $\mathbb{G}_{\underline{\alpha}_t}$ admettant une fonction de masse g. On suppose que $n = mS$, $m \in \mathbb{N}^*$ alors on a d'après (Gladyshev (1963)) $t = s + \tau S, s = 1, \ldots, S$, avec $\tau = 0, 1, \ldots, m-1$.

3.1 Estimation des paramètres du modèle

Proposition 3.1. (*Estimateur des moindres carrées conditionnels*). *Soit* $\underline{\varphi}_s \in [0,1]^p$, $\mathbb{G}_{\underline{\alpha}_s}$ tel que $\mathbb{E}_{\mathbb{G}_{\underline{\alpha}_s}}\left[\varepsilon_0\right]^3 < \infty$ *et* $g_{\underline{\alpha}_s}(0) \in (0,1)$ *alors* :$\overline{\underline{\varphi}}_{s,m} \xrightarrow{\mathcal{L}} \underline{\varphi}_s$ and $\overline{\mu}_{\underline{\alpha}_s,m} \xrightarrow{\mathcal{L}} \mu_{\mathbb{G}_{\underline{\alpha}_s}}$[2]

$$\begin{pmatrix} \overline{\underline{\varphi}}_{s,m} \\ \overline{\mu}_{\mathbb{G}_{\underline{\alpha}_s},m} \end{pmatrix} = \sum_{\tau=0}^{m-1} \begin{pmatrix} y_{(s-1)+\tau S}^2 & \cdots & y_{(s-1)+\tau S} \\ y_{(s-2)+\tau S} y_{(s-1)+\tau S} & \cdots & y_{(s-2)+\tau S} \\ \vdots & \ddots & \vdots \\ y_{(s-1)+\tau S} & \cdots & m \end{pmatrix}^{-1} \times \begin{pmatrix} \sum_{\tau=0}^{m-1} y_{(s-1)+\tau S} y_{s+\tau S} \\ \sum_{\tau=0}^{m-1} y_{(s-2)+\tau S} y_{s+\tau S} \\ \vdots \\ \sum_{\tau=0}^{m-1} y_{s+\tau S} \end{pmatrix}$$

Proposition 3.2. (*Estimateur du maximum de vraisemblance conditionnel*). *Un estimateur* $\left(\widehat{\underline{\theta}}_n\right)_{n\in\mathbb{Z}_+}$ *de* $\underline{\theta}$ *est un estimateur MVC de* $\underline{\theta}$, *si* $\widehat{\underline{\theta}}_n$ *maximise la vraisemblance conditionnelle du modèle, i.e.* $\widehat{\underline{\theta}}_{s,m} = \left(\widehat{\underline{\varphi}}_{s,m}, \widehat{\underline{\alpha}}_{s,m}\right)$ *maximise les s vraisemblances si et seulement si les conditions suivantes sont retenues* :

a) $\widehat{g}_{\underline{\alpha}_s,m}(e) = 0$ *pour* $e < 0$ *et* $e > u_{s+}$, *avec* $u_{s+} = \max_{\tau=0,\ldots,m-1}(y_{s+\tau S})$

b) $\left(\widehat{\underline{\varphi}}_{s,m}, \widehat{\underline{\alpha}}_{s,m}\right)$ *est une solution pour le problème d'optimisation polynomial contraints*

$$\max_{\substack{x_{s,1},\ldots,x_{s,p} \\ z_{s,1},\ldots,z_{sq}}} \left\{ \prod_{\tau=0}^{m-1} \left(\sum_{e=0}^{y_{s+\tau S}} g_{\underline{z}_s}(e) \sum_{\substack{0\le k_l \le y_{s-l+\tau S}, l=1,\ldots,p \\ k_1+\ldots+k_p = y_{s+\tau S}-e}} \left[\prod_{l=1}^{p} \binom{y_{(s-l)+\tau S}}{k_l} x_{s,l}^{k_l} (1-x_{s,l})^{y_{(s-l)+\tau S}-k_l} \right] \right) \right\}$$

sous les contraintes

$0 \le x_{s,i} \le 1$ *pour* $s \in \{1,\ldots,S\}$ *fixé et* $i = 1,\ldots,p$,
$g_{z_{s,j}} \ge 0$ *pour* $s \in \{1,\ldots,S\}$ *fixé et* $\forall j = 1,\ldots,q$ avec $\sum_{e=0}^{u_{s+}} g_{z_{s,j}}(e) = 1$.

Nous soulignons que nous n'imposerons nulle part qu'un tel maximum local est unique.

3.2 Prédiction à court-terme du modèle

Soit $\mathcal{F}_n = \sigma(y_1,\ldots,y_n)$ la σ-algèbre générée par $y_1, y_2, \ldots, y_n$, alors le prédicteur de variance minimale $y_n(1)$ de y_{n+1}est donné par

$$\widehat{y}_n(1) = \mathbb{E}_{\underline{\theta}}(y_{n+1}|\mathcal{F}_n) = \sum_{i=1}^{p} \varphi_{1,i} y_{n+1-i} + \mu_{\mathbb{G}_{\underline{\alpha}_1}},$$

nous pouvons constater que pour $h = 2$ le prédicteur de variance minimale $y_n(2)$ de y_{n+2} est donné par $\widehat{y}_n(2) = \mathbb{E}_{\underline{\theta}}(y_{n+2}|\mathcal{F}_n) = \sum_{i=1}^{p} \mathbb{E}_{\underline{\theta}}\left(\varphi_{2,i} \circ y_{n+2-i}|\mathcal{F}_n\right) + \mu_{\mathbb{G}_{\underline{\alpha}_2}}$,

et on a d'aprés (Yan, 1985)

$$\begin{aligned} \mathbb{E}_{\underline{\theta}}\left(\varphi_{2,i} \circ y_{n+2-i}|\mathcal{F}_n\right) &= \mathbb{E}_{\underline{\theta}}\left(\mathbb{E}_{\underline{\theta}}\left(\varphi_{2,i} \circ y_{n+2-i}|y_{n+2-i}, \mathcal{F}_n\right)|\mathcal{F}_n\right) \\ &= \varphi_{2,i}\mathbb{E}_{\underline{\theta}}(y_{n+2-i}|\mathcal{F}_n) = \varphi_{2,i}\widehat{y}_{n+2-i} = \varphi_{2,i}\widehat{y}_n(2-i) \end{aligned}$$

par induction nous pouvons donner la formule générale pour un $h > 2$ en tenant compte

[2]La notation $X \xrightarrow{\mathcal{L}} Y$ désigne la convergence en loi de la variable aléatoire X.vers Y

de la périodicité des paramètres du modèle, sous la proposition ci-dessous

Proposition 3.3. (*Prédicteur à court terme de variance minimale*). *Soit* $\mathcal{F}_n = \sigma(y_1, \ldots, y_n)$ *la* σ*-algèbre générée par* $y_1, \ldots, y_n$*, alors le prédicteur de variance minimale* $y_n(h)$ *de* y_{n+h} *est donné pour* $h > 2$ *par*

$$\widehat{y}_n(h) = \mathbb{E}_{\underline{\theta}}(y_{n+h}|\mathcal{F}_n) = \sum_{i=1}^{p} \varphi_{r+S,i}\widehat{y}_n(h-i) + \mu_{\mathbb{G}_{\underline{\alpha}_{r+S}}},$$

où r *est le reste de la division euclidienne de* h *sur* S*, c-à-d,* h *et* r *sont congruents modulo* S *et on ecrit* $h \equiv r\,[S]$.

4 Exemple Numérique

4.1 Etude de simulation

Dans cette sous-section, nous avons évalué les estimateurs du (MCC) et ceux du (MVC), sur une série chronologique générée à partir d'un modèle $PINAR_4(2)$, pour différentes tailles, $(n = 80, 400, 1000, 2000)$. Nous avons considérés un modèle $PINAR_4(2)$, où la distribution d'innovation suit une distribution Géométrique $\mathcal{G}(e^{-\alpha_s})$, $s = 1, 2, 3, 4$, :

$$\begin{aligned}\text{Modèle} : \underline{\theta} &= \left[\left(\varphi_{1,1}, \varphi_{1,2}, \exp(-\alpha_1)\right); \ldots; \left(\varphi_{4,1}, \varphi_{4,2}, \exp(-\alpha_4)\right)\right]', \\ &= \left[(.15, .45, \exp(-4)); (.50, .2, \exp(-1)); (.76, .38, \exp(-3.5)); (.63, .93, \exp(-2))\right]'.\end{aligned}$$

Table 1. *Résultats de simulation concernant le Model* $PINAR_4(2)\,|\mathcal{G}(e^{-\alpha_s})$

	s	$\varphi_{s,1}$	$\widehat{\varphi}_{s,1}$	$RMSE_s$	$\varphi_{s,2}$	$\widehat{\varphi}_{s,2}$	$RMSE_s$	$e^{-\alpha_s}$	$\widehat{e^{-\alpha_s}}$	$RMSE_s$
MCC $n=$ 1000	1	.15	.1637	.1064	.45	.4525	.1696	.0183	.0176	.0791
	2	.50	.4995	.0129	.20	.1999	.0097	.3678	.3371	.6336
	3	.76	.7450	.1219	.38	.3865	.0870	.0301	.0300	.0586
	4	.63	.6295	.0209	.93	.9307	.0288	.1353	.1568	.0958
MVC $n=$ 1000	1	.15	.1609	.0794	.45	.4659	.1296	.0183	.0185	.0034
	2	.50	.5005	.0107	.20	.1997	.0070	.3678	.3635	.4051
	3	.76	.7575	.0662	.38	.3835	.0344	.0301	.0302	.0051
	4	.63	.6295	.0114	.93	.9302	.0222	.1353	.1563	.0704

La table 1 représente une partie des résultats de la simulation pour 1000 réplications concernant la taille 1000, où les estimations sont très proches de leurs vraies valeurs selon le critère du $RMSE$, ce qui implique la convergence des deux méthodes d'estimations.

4.2 Etude sur données réelles

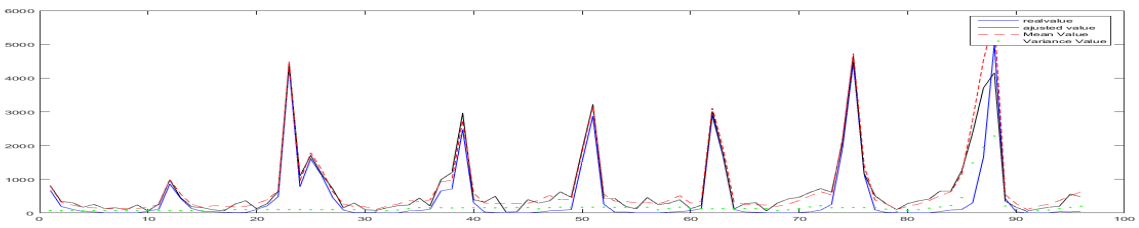

Figure 1. *différentes trajectoires ajustées du processus* y_t *opposées aux vraies données*

Dans cette sous-section, nous considérons une série de données saisonnière composée de 96 observations, cette dernière représente le nombre mensuel de personnes diagnostiqués en état de grippe au niveau des services hospitaliers de la région de la Catalogne (Espagne) entre 2009 et 2016. Cette série chronologique est illustrée dans la figure 1, ainsi que différentes trajectoires ajustées du processus y_t générées à partir d'un un modèle $PINAR_{12}(2)$ conduit par une distribution marginale géométrique.[3] Pour la prévision du nombre d'admissions hospitalières, nous avons illustrés la qualité de notre prédicteur à court-terme de variance minimale dans la figure 2. Nous avons établis le graphique de comparaison entre les 12 dernières vraies valeurs et leurs valeurs prédites, et cela après avoir enlevé 12 observations de la série mensuelle originale de taille 96, donc nous avons utilisés 84 observations pour prédire le nombre mensuel d'admissions durant l'année 2016.

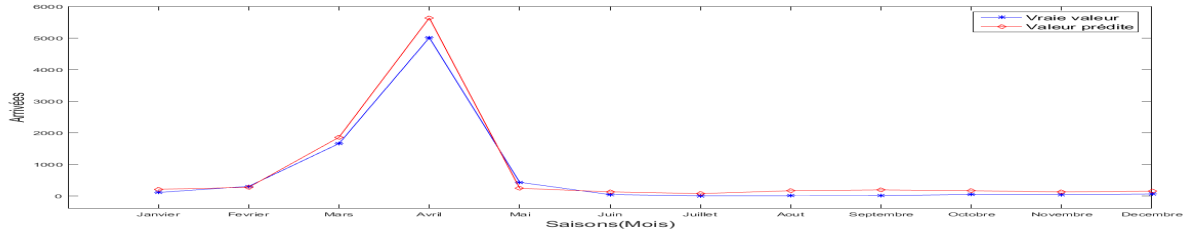

Figure 2. Comparaison entre les 12 dernières vraies valeurs et leurs valeurs prédites

Bibliographie

Alzaid, A.A. and Al-Osh, M. (1990). Integer-valued p th-order autoregressive structure ($INAR(p)$) process. *Journal of Applied Probability* 27 pp. $314-324$.

Bourguignon, M. Vasconcellos, K L. P. Reisen, V. A. and Ispany, M. (2016). A Poisson $INAR(1)$ process with a sesonal structure. *Journal of Statistical Computation and Simulation.* 86 (2) pp. $373-387$.

Du, J-G and Li, Y. (1991). The integer-valued autoregressive ($INAR(p)$) model. *Journal of Time Series Analysis.* 12(2) pp. $129-142$.

Gladyshev, E.G. (1963). Periodically and Almost-Periodically Correlated Random Processes With Continuous Time Parameter. *Theory Probab & its App.* 8 (2) pp. $173-177$.

Moriña, D. Puig, P. Rios, J. Viella, A. and Trilla, A (2011). A statistical model for hospital admissions caused by seasonal diseases. *Journal of statistics in Medecine.* 30 (26) pp. $3125-3136$.

Rao. Y. and McCabe. B. (2016). Real Time Monitoring for Abnormal Events Pattern: an Application to influenza Outbreaks. *Statistics in Medicine.* 35 (13) pp. $2206-2220$.

Steutel, F. W. and Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability,* Vol. 7, No. 5, $893-899$.

Yan. Shi-Jian. (1985). *Probability Theory*, Bejing: Scientific Theory.

[3]Moyenne et variance de 1000 trajectoires d'un $PINAR_{12}(2)\,|\mathcal{G}(e^{-\alpha_t})$ sont illustrées dans la figure 1